大数据看两会
——百分点数据新闻的探索

本刊记者｜陈旭管　编辑整理

编者按：今年“两会”期间，民生、互联网、文化、创业、环保、人工智能等热门议题广受关注。两会新闻滚动刷屏，哪个话题被提及最多？如何从中读懂中国未来？大数据有奇招！

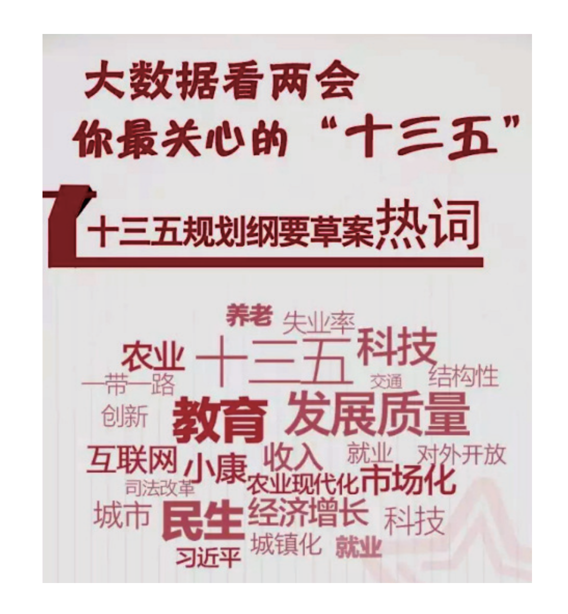
数据新闻报道分为四个环节：对原始数据的收集，对数据的分析和过滤，将数据可视化，形成具体的新闻报道。在数据收集方面，百分点通过舆情系统，全面覆盖网媒，及时抓取信息，准确分析语义，配合舆情分析模块，深度理解群众对热点的讨论态度。其中网络舆情数据抓取方法如下：

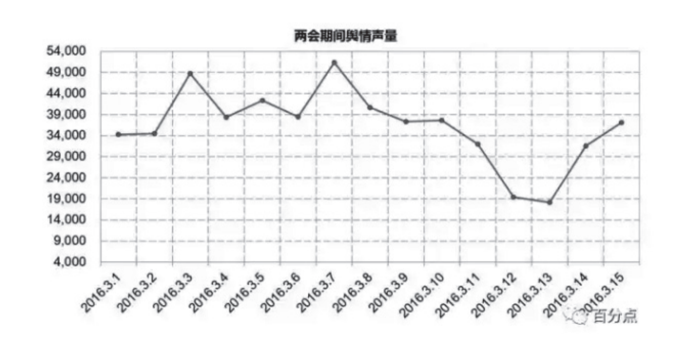
监测方法：相关的新闻、论坛、贴吧、微博、微信历史数据在限定的时间段内进行爬取；通过文本分词、数据预处理，利用 tfidf 模型计算每个词语的权重，形成词语的热度；对相关信息进行聚类或分类并判断其正面、中性或负面情感。

监测周期：过去半年。

统计指标：声量（关注度）、话题热度、话题情感等。

最后形成数据可视化报告，分析两会热词和群众讨论的声量走向。







从 2016 年起，百分点公司支撑新华社有关部门关于“两会”大数据新闻报道的需求，依托百分点公司的舆情系统和文本分析系统，为新华社有关部门提供稳定、可靠的大数据技术支撑，开展“大数据看两会”系列专题建设。系列报道及可视化产品在新华社“两微一端”平台落地。

在 2017 年新华社与百分点合作的“大数据看两会”新闻报道中，主要盘点了 2014 至 2017 年两会热点话题，并针对以下维度做了分析：

①历年两会期间的话题关注榜；②历年男性、女性、不同年龄、不同职业群体的关注点，以及所属地域划分；③历年两会期间点赞（浏览）数最多的报道，被哪些群体热议？

2017 年两会，百分点公司还与人民日报中央厨房、中国日报等媒体合作推出专题。

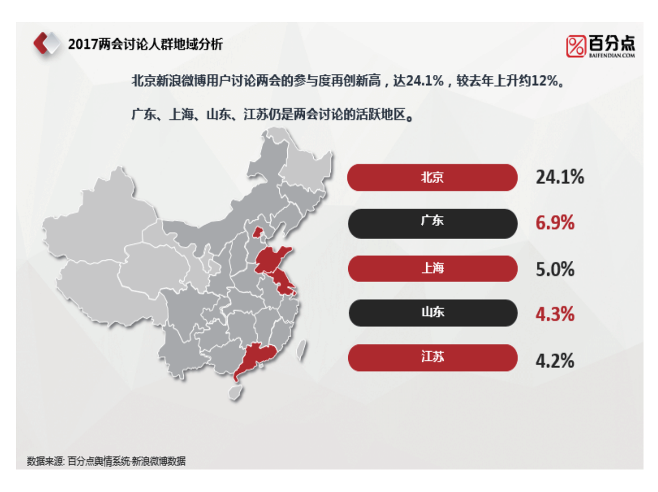
百分点舆情系统监测于 2017 年 3 月 1 日至 9 日对新华网、

人民网、网易、新浪网、今日头条、搜狐网、新浪微博等全网数据，进行分析解读，通过解读发现如下舆情趋势：

①发展是根本，也是两会最重要的话题。



②北京新浪微博用户讨论两会的参与度再创新高，达24.1%，较2016年上升约12%。广东、上海、山东、江苏仍是两会讨论的活跃地区。



③分析各地区人群对两会热点关注的差异。

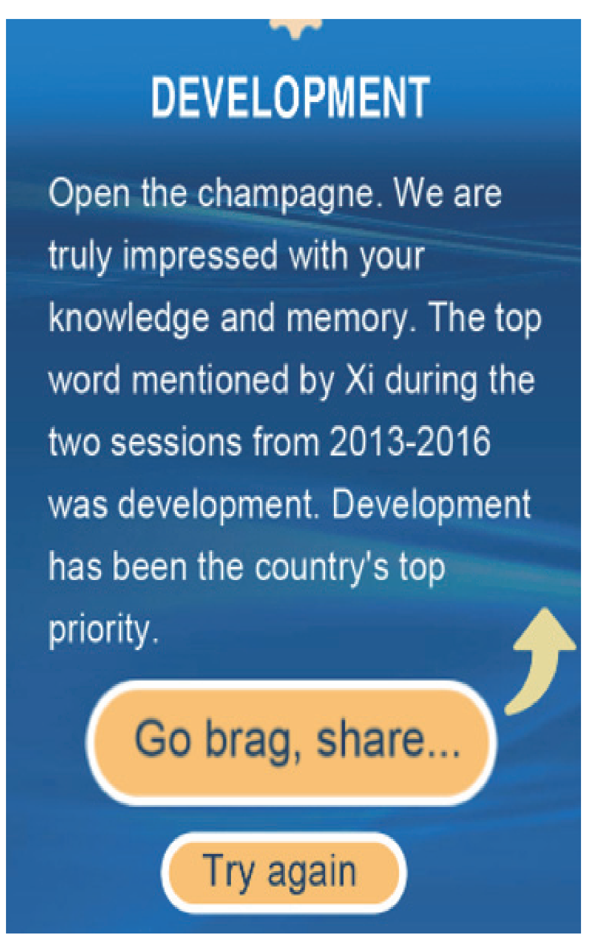


2017两会讨论人群地域关注差异

	省市	关注话题		省市	关注话题
1	北京	民生、互联网、环保	11	河北	民生、互联网、文化产业
2	广东	创业、互联网、环保	12	湖北	文化产业、国企改革、经济发展
3	上海	创业、经济发展、国企改革	13	陕西	互联网、民生、外交
4	山东	环保、互联网、民生	14	辽宁	国企改革、互联网、民生
5	江苏	民生、互联网、文化产业	15	安徽	民生、互联网、反腐
6	浙江	创业、外交、科技文化	16	山西	民生、社会治安、房地产
7	福建	民生、住房、互联网	17	天津	供给侧改革、社会治安、民生
8	河南	民生、创业、互联网	18	黑龙江	民生、互联网、房地产
9	四川	互联网、民生、反腐	19	江西	民生、互联网、房地产
10	湖南	创业、科技文化、环保	20	重庆	创业、养老改革、民生

数据来源: 百分点舆情系统·新浪微博数据







党和国家领导人参加团组审议和讨论的情况，是外界重点关注的话题。在与中国日报合作的两会专题中，百分点公司首先根据相关议题关键字进行数据收集，其次通过运用模型算法，文本分析系统识别非结构化文本，还针对领导人参加团组会议的舆情进行情感分析，话题和口碑检测等业务。

百分点公司与人民日报社、新华社、中国日报等央媒的合作主要包括两会前收集数据，预测热点话题，到通过舆情系统、模型算法、文本分析系统支撑两会专题报道。从这些合作中可以看出数据新闻已经成为媒体行业的迫切需求，也是媒体转型发展的趋势。对媒体机构而言，大数据技术已经渗透到新闻报道的各个环节，包括：挖掘新闻热点；追踪事件提升新闻报道质量；利用数据可视化表达创新新闻报道形态；实施精准用户画像；量化建立媒体平台的传播力指标并评估传播效果；利用分布式计算、机器学习等多种手段，分析用户使用特点和规律等等。媒

SCI-TECH

A good attempt. China has planted its footprints as high as the outer space and as low as the bottom of oceans. Xi urged experts to achieve breakthroughs in key technology areas. But it was not the top word. Have another try.

Try again

Share